CS 7545: Machine Learning Theory **Fall 2021**

Lecture 13

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 VC-Dimension

We saw that our previous algorithms have a $\frac{1}{\gamma^2}$ dependence on the margin; however, this can in fact be exponential (*e.g.*, in decision lists), so we prefer a $\log \frac{1}{\gamma}$ type dependence. This is indeed possible, but with an n in the numerator rather than the constant or $\log n$ of winnow and perceptron.

Suppose we are given data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell$. We are interested in the set of feasible hypotheses $W = \{\boldsymbol{w} : \|\boldsymbol{w}\|_2 = 1 \text{ and } \operatorname{sgn}(\boldsymbol{w}^\top \boldsymbol{x}_i) = y_i\}$. Upon receiving $\boldsymbol{x}_{\ell+1}$, we look at the subsets

$$W \cap \{\boldsymbol{w} : \|\boldsymbol{w}\|_2 = 1 \text{ and } \boldsymbol{w}^\top \boldsymbol{x}_{\ell+1} \geq 0\} \tag{1}$$

$$W \cap \{\boldsymbol{w} : \|\boldsymbol{w}\|_2 = 1 \text{ and } \boldsymbol{w}^\top \boldsymbol{x}_{\ell+1} < 0\}; \tag{2}$$

the majority algorithm will choose the subset with the largest volume (we are just selecting the most likely label). So, after n mistakes, $\operatorname{vol}(W) \leq \frac{1}{2^m}\operatorname{vol}(B_n)$.

Recall $\gamma = \min_{\boldsymbol{x}} |\boldsymbol{w}^\top \boldsymbol{x}|$. An alternative interpretation of the margin is that it is the set of hypotheses that will always be feasible. That is, any $\boldsymbol{w}$ in the γ-cone around $\boldsymbol{w}^\star$ will label every example correctly. So, we will never eliminate the γ-cone from W. Taking integrals, we find

$$\frac{\operatorname{vol}(\gamma\text{-cone})}{\operatorname{vol}(B_n)} = c\gamma^n. \tag{3}$$

So,

$$\frac{1}{2^m} \geq c\gamma^n \implies m \leq cn \log \frac{1}{\gamma} \tag{4}$$

as desired.

How do we actually calculate the volume of the feasible subsets? Well, since we can sample from any convex body efficiently, we can just use random sampling to estimate the volume fraction. In fact, if we choose a single point, there is at least a $\frac{1}{2}$ probability that it lies in the majority set, so such a technique will at most double our mistake bound.

Instead of dealing with the feasible subsets, we could also just keep track of a single feasible hyperplane. This motivates the general setup of PAC-learning: if $h^\star \in \mathcal{H}$ and $x_1, x_2, \ldots, x_\ell \sim D$ with labels, we can find an empirical risk minimizer (ERM) $h \in \mathcal{H}$ such that $h(x_i) = h^\star(x_i)$ for all $i = 1, 2, \ldots, \ell$. We'd like to know how many samples m we need for h to be probably approximately correct:

$$\Pr_{x \sim D}[h(x) \neq h^\star(x)] \leq \epsilon \text{ with probability} \geq 1 - \delta. \tag{5}$$

Presumably, m is a function of ϵ, δ, and the complexity of $\mathcal{H}$. But the complexity is not the same as the size of $\mathcal{H}$ – it is more closely related to the number $\mathcal{H}[m]$ of distinct ways concepts in $\mathcal{H}$ can label a set S of m points.

Definition 1.1: VC-dimension

VCdim($\mathcal{H}$) is the largest integer m such that there exists a set of m points which can be shattered (*i.e.*, labeled in all 2^m ways) by concepts in $\mathcal{H}$.

For example, intervals on a line have VC-dimension 2, axis-parallel rectangles have VC-dimension 4, and halfspaces in d dimensions have VC-dimension $d+1$. Note that, to show $\mathrm{VCdim}(\mathcal{H}) = m$, we need to show that there exists a set of m shatterable points, and that all sets of size $m+1$ are not shatterable.

Lemma 1.1: Sauer-Shelah Lemma

Suppose $\mathrm{VCdim}(\mathcal{H}) = d$. Then, $\mathcal{H}[m] \leq \sum_{i=1}^{d} \binom{m}{i} = \binom{m}{\leq d} \leq m^d$.

Theorem 1.1: PAC-learning

The number of examples needed to (ϵ, δ)-PAC learn $\mathcal{H}$ is

$$m \leq \frac{2}{\epsilon}\left(\log 2\mathcal{H}[2m] + \log \frac{1}{\delta}\right). \tag{6}$$

Corollary 1.1: VC-dimension bound on sample complexity

The number of examples needed to (ϵ, δ)-PAC learn $\mathcal{H}$ is

$$m \leq \mathcal{O}\left(\frac{1}{\epsilon}\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right). \tag{7}$$

What if realizability is false (*i.e.*, there is no perfect hypothesis $h^\star$)? Then, we want to be as close as possible to the best hypothesis in the class:

$$\mathrm{OPT}_{\mathcal{H}} = \min_{h \in \mathcal{H}} \Pr_{x \sim D}[h(x) \neq \ell(x)]. \tag{8}$$

Let

$$\mathrm{err}_S(h) = \frac{|\{x \in S : h(x) \neq \ell(x)\}|}{|S|}. \tag{9}$$

Then, we want to show that on the total distribution,

$$\mathrm{err}_D(h) \leq \mathrm{OPT}_{\mathcal{H}} + \epsilon. \tag{10}$$

The following theorem is much more general. It shows that we closely approximate the errors of *all* hypotheses, and approximating the best hypothesis is a special case.

Theorem 1.2: Uniform Convergence

Suppose the number of examples is

$$m \geq \frac{8}{\epsilon^2}\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right). \tag{11}$$

Then, with probability at least $1 - \delta$, for all $h \in \mathcal{H}$,

$$|\mathrm{err}_D(h) - \mathrm{err}_S(h)| \leq \epsilon. \tag{12}$$

We will now show the proofs. For the lemma, we will use the following fact.

Fact 1.1

$\binom{m}{\leq d} = \binom{m-1}{\leq d} + \binom{m-1}{\leq d-1}$.

Proof: (of Lemma 1.1). Suppose S is a set of m points drawn from D. We will perform induction on m, and the base case $m \leq d$ is obvious.

Suppose $x \in S$ and consider $S \setminus \{x\}$. By the inductive hypothesis,

$$\mathcal{H}[S \setminus \{x\}] = \mathcal{H}[m-1] \leq \binom{m-1}{\leq d}. \tag{13}$$

By the fact, it suffices to show that

$$\mathcal{H}[S] - \mathcal{H}[S \setminus \{x\}] \leq \binom{m-1}{\leq d-1}. \tag{14}$$

For a labeling to be in $\mathcal{H}[S]$ but not in $\mathcal{H}[S \setminus \{x\}]$, there must be labelings h, h' which agree on all points except x. Define the set of these labelings

$$T = \{h \in \mathcal{H}[S] : h(x) = 1, h'(x) \in \mathcal{H}[S]\}. \tag{15}$$

Let $\mathrm{VCdim}(T) = d'$, so there exists a set of $2^{d'}$ points shattered by T. But then $d' + 1$ points can be shattered by $\mathcal{H}$. So, $d' + 1 \leq d$, so $\mathrm{VCdim}(T) \leq d - 1$. Therefore,

$$\mathcal{H}[T] \leq \binom{m-1}{\leq d-1}. \tag{16}$$

■